CALCULATING VARIANCES USING DATA FROM THE HCUP NATIONWIDE INPATIENT SAMPLE, 1988-1997

Sarah Q. Duffy, Ph.D. and John P. Sommers, Ph.D.
Agency for Health Care Policy and Research

The principals outlined in this document will apply to the 1998 NIS as well. A revised version of this document will be made available in the near future and will apply specifically to the new file structure beginning with the 1998 NIS.

INTRODUCTION

The Nationwide Inpatient Sample (NIS), Release 1 database contains all discharges from hospitals that were selected without replacement according to a stratified probability sample design from a frame that includes hospitals from 8 states for 1988 and 11 states for 1989-1992. Release 2 and Release 3 include data from 17 states for 1993 and 1994, respectively, and Release 4 and Release 5 include data from 19 states for 1995 and 1996, and Release 6 includes data from 22 states in 1997. Failure to account for this sample design when computing statistics will cause variances to be estimated incorrectly. This document states the problem and gives an example of how one readily available complex survey design package, the Survey Data Analysis Software System, or SUDAAN, can be used to estimate variances while accounting for the sample design of the NIS. The reader should be prepared to consult the SUDAAN documentation as necessary.

Due to the correlation between observations caused by the same hospitals appearing in the NIS data across years, it is difficult to calculate standard errors when multiple years of data are pooled in one analytic dataset. The methods described in this paper are appropriate for calculating variances using one year of NIS data at a time.

The principals outlined in this document will apply to the 1998 NIS as well. A revised version of this document will be made available in the near future and will apply specifically to the new file structure beginning with the 1998 NIS.

BACKGROUND

Variances Based on Simple Random Sampling

Many popular statistical packages, such as SAS and SPSS, use the following formula based on simple random sampling to calculate an estimate for the sample variance:

$$\hat{\sigma}^2 = \frac{\sum (y_{hij} - \bar{y})^2}{n - 1}$$

where:

$\hat{\sigma}^2$ = variance estimate

σ = the standard deviation

y_{hij} = the value of variable *y* for the *jth* sample discharge in the *ith* sample hospital in the *hth* stratum

$\bar{y}$ = the grand mean of the variable y, and

n = the number of observations in the sample.

Variances Based on the NIS

Since the NIS is not a simple random sample, it requires a different variance formula. The NIS sample design has several characteristics that require modification of the variance formula: sample weights, two-stage sampling from a finite population, and stratification. Complex survey design packages such as SUDAAN (descriptive statistics), SURREGER (ordinary least squares regression), and RTILOGIT (logistic regression) allow these characteristics to be incorporated into variance estimation.[1]

Variance formulas appropriate for the NIS data contain weights, components for the two stages of sampling, and factors to correct for the proportion of the frame included in the sample at each level (finite population correction factors). For example, define the weighted sum, Y,

$$Y = \sum_{h}^{H} \sum_{i}^{n_h} \sum_{j}^{n_{hi}} w_{hij} y_{hij}$$

where:

y_{hij} = the value of a variable *y* for the *jth* sample discharge in the *ith* sample hospital in the *hth* stratum, as above

w_{hij} = a set of weights or any other constants over the set of sample discharges, hospitals, and strata

n_{hi} = the number of discharges in the *ith* sample hospital in the *hth* stratum

n_h = the number of sample hospitals in the *hth* stratum and

H = the number of strata.

Then the estimate of the variance of Y from the sample, $\hat{\sigma}_Y{}^2$, is

$$\hat{\sigma}_Y^2 = \sum_{h}^{H} (1-f_h) n_h S_h^2 + \sum_{h}^{H} f_h \sum_{i}^{n_{hi}} (1-f_{hi}) n_{hi} S_{hi}^2 \tag{1}$$

where:

$\hat{\sigma}$ = the standard deviation of Y

f_h = the proportion of the total number of hospitals in the *hth* stratum selected into the sample, i.e., the first stage sampling rate in the *hth* stratum. (This is simply the number of hospitals from stratum h in the sample divided by the total number of hospitals in stratum h on the frame.)

f_{hi} = the proportion of the discharges in the sample from the *ith* sample hospital in the *hth* stratum, i.e., the second stage sampling rate in the *hith* hospital.[2] (This is simply the number of discharges from hospital *i* in stratum *h* in the sample divided by the total number of discharges in hospital *i* and stratum *h.*)

S_h^2 = the component for the first stage of sampling, the overall variation due to variation between hospitals within strata

$$= \frac{\sum_{i}^{n_h}\left(\sum_{j}^{n_{hi}} w_{hij}y_{hij} - \frac{\sum_{i}^{n_h}\sum_{j}^{n_{hi}} w_{hij}y_{hij}}{n_h}\right)^2}{(n_h-1)} \text{ and}$$

$S_{hi}{}^2$ = a portion of the component for the second stage of sampling, the overall variation of discharges within hospital for the *hith* hospital.

$$= \frac{\sum_{j}^{n_{hi}}\left(w_{hij}y_{hij} - \frac{\sum_{j}^{n_{hi}} w_{hij}y_{hij}}{n_{hi}}\right)^2}{n_{hi}-1} \; .$$

SUDAAN uses variance formulas similar to (1) in its DESCRIPT procedure to calculate a large number of simple statistics, such as estimates of means and totals. Variances of more complex estimates, such as ratios of two different random variables, require formulas that include contributions from the variances of each of the random variables. These can be calculated using the SUDAAN RATIO procedure.

Because the states included in the NIS frame were not selected randomly, the error associated with statistical estimates derived from NIS data actually contains two parts:

a. variance formula (1), the error due to sampling from the selected states, and
b. the bias from not using the entire U.S. as the sampling frame.

Part b, the bias, cannot be calculated directly. When possible, estimates based on the NIS should be checked against national benchmarks, such as Medicare data or data from the National Hospital Discharge Survey, to determine the appropriateness of the NIS for specific analyses.

AN EXAMPLE: CALCULATING MEAN LENGTH OF STAY FOR AMI PATIENTS

SUDAAN, available for both the PC and the mainframe, can be used to calculate variance estimates for simple descriptive statistics (means, percentiles, ratios) and cross tabulations.[3] For example, suppose an analyst wants to calculate a nationally weighted mean length of stay (LOS) and its variance for all patients in the NIS with acute myocardial infarction (AMI) as a principal diagnosis, using the SUDAAN DESCRIPT procedure.

Description of SUDAAN Code

The SUDAAN code would be as follows:

```
PROC DESCRIPT  DATA=analysis file  FILETYPE=analysis file type  DESIGN=WOR;

WEIGHT DISCWT_U;

NEST STRATUM HOSPID/MISSUNIT;

TOTCNT N_HOSP_F _ZERO_;

SAMCNT S_HOSP_U TOTDSCHG;

VAR LOS;

PRINT MEAN SEMEAN;

OUTPUT MEAN SEMEAN/ FILETYPE = ASCII FILENAME = "output file";
```

Details:

PROC DESCRIPT

DATA = the name of the analysis file. In this example, the analysis file consists of all discharges in the NIS with an AMI coded as a principal diagnosis. The input analysis file must always be sorted by the variables specified in the NEST statement (see below), in this case STRATUM and HOSPID.

FILETYPE = the format of the analysis file. SUDAAN will read SAS and ASCII files.

DESIGN = the sample design. For the NIS, WOR (without replacement) is the appropriate choice. See the SUDAAN documentation for details.

Requests for statistics other than means, such as quantiles, would be included in the PROC DESCRIPT statement as well. Since mean is the default statistic, it is not necessary to specify it.

WEIGHT

Specifies the weighting variable. It is a required statement. The NIS variable DISCWT_U is the weight for discharges in the NIS, and must be merged onto the analysis file from the NIS, Release 1 Hospital Weights file. (DISCWT_U has been merged to Inpatient Stay Core File A for NIS, Release 2, Release 3, Release 4 and Release 5.) To get unweighted statistics, simply create a variable that is equal to 1 for all observations and specify that variable in the WEIGHT statement.

NEST

Specifies the variables corresponding to the levels in the sampling design. It is a required statement. In the NIS, hospitals were selected from strata, which are identified by the NIS variable STRATUM, and discharges were selected from hospitals, which are identified by the NIS variable HOSPID. The analysis file must be sorted by the variables in this statement.

TOTCNT

Specifies the population counts at each stage of the sampling design for which without replacement sampling is assumed. It is a required statement. The NIS variable N_HOSP_F contains the number of hospitals on the frame in the stratum, and must be merged onto the analysis file from the NIS hospital-level file.[4]

ZERO is a variable available to all SUDAAN procedures that is used to identify stages for which no sample selection took place. It prompts SUDAAN not to calculate the corresponding variance component. It is used in this example for the hospital-level counts because there was no sampling at this level - all AMI discharges are included in the analysis file.[5] This makes each f_{hi} equal to one in formula (1) effectively zeroing the second term. Hence, the name _ZERO_.

SAMCNT

Specifies the sample counts at each stage of the sample design for which without replacement sampling is assumed. It is an optional statement. The NIS variable S_HOSP_U, again from the hospital file, contains the number of hospitals sampled in each stratum. The NIS variable TOTDSCHG contains the number of discharges in the hospital, which must be specified even though _ZERO_ is specified.

SUDAAN double checks the accuracy of the variables specified in the SAMCNT statement by counting the number of observations in the file at each level. It will do this whether or not the sample count variables are specified. In this case, since SUDAAN's count will be the correct value, it may make most sense to omit the SAMCNT statement from program. However, there are applications for which SUDAAN's count will be incorrect. See the section "Using the NIS Subsamples", below, for an example of such an application.

VAR

Specifies the variables for which statistics are to be calculated, in this case LOS. Multiple variables may be included on the VAR statement, but they must all be of one type, continuous or discrete.

PRINT

Specifies that the mean (MEAN) and its standard error (SEMEAN) be printed. SEMEAN, the standard error of the mean, is equivalent to the $\hat{\sigma}$ referred to earlier and should be used when calculating Z scores and other tests of significance.

OUTPUT

Specifies the ASCII output files to which the results are to be read and requests that MEAN and SEMEAN be included on the files. SUDAAN creates 5 files, each of which has as its root name the name specified on the statement. Details may be found in the SUDAAN documentation.

Steps in Computation

To demonstrate the effect of using SUDAAN, standard errors were calculated using both the above program and SAS PROC MEANS. This involved:

1. Pulling all discharges from the 1992 NIS with DCCHPR1 = 100 (acute myocardial infarction).

2. Merging the resulting file to the NIS, Release 1 Hospital Weights File.

3. Downloading the data to a secure PC environment.

4. Running PC SUDAAN PROC DESCRIPT to get weighted national estimates of

   a. the number of AMI patients, and
   b. mean length of stay of AMI patients and its standard error.

5. Running PC SAS to get both weighted and unweighted estimates of the variables mentioned in step 4. The SAS weighted estimates were computed two ways:

   a. using the WEIGHT Statement with the VARDEF=WEIGHT option. When using the WEIGHT statement to get counts, the analyst must specify a variable equal to 1 for each discharge and request a SUM for that variable.
   b. using the FREQ statement. Since the weight variables in the NIS are not integers, using the FREQ statement will underestimate counts, as the results below reveal. This is because SAS uses only the integer portion of variables specified in the FREQ statement.

Results of Computation

The results, displayed in Table 1, reveal that accounting for the sample design affects the estimated variances. In this example, accounting for the sample design resulted in lower variances. The standard error calculated by SUDAAN is only 50% as large as the one calculated using SAS PROC MEANS with the FREQ statement, which as noted above gives incorrect count estimates as well, and less than one quarter the size of that estimated when using SAS PROC MEANS either unweighted or with the WEIGHT statement. Calculating the variances using SUDAAN or other complex survey design package will often result in lower variances from NIS data because of the finite correction factor, but this is by no means guaranteed.

Table 1 Weighted and Unweighted Estimates of Counts, and Average Length of Stay and its Standard Error for 1992 AMI Discharges, HCUP National Inpatient Sample, Release 1

Variable	Weighted PC SUDAAN	Weighted PC SAS Weight Statement[2]	Weighted PC SAS, FREQ Statement[3]	Unweighted PC SAS
Count of AMIs[1]	688,054	688,054	625,605	119,121
Average Length of Stay (Standard Error)	7.88 (0.05)	7.88 (0.21)	7.87 (0.10)	8.01 (0.23)

Notes:
1. AMI discharges are those with DCCHPR1 = 100.
2. Run with the VARDEF = WEIGHT statement. When a WEIGHT statement is specified, the reported N is equal to the unweighted sum. To find the weighted sum the analyst must create a variable that equals 1 for each observation and request SUM on that variable.
3. PROC FREQ results in a lower weighted count because it only uses the integer portion of the weights. As the results reveal, it should not be used to compute weighted estimates from the NIS data.

USING THE NIS SUBSAMPLES

SUDAAN can estimate variances using data from a NIS 10% subsample, but the program must be modified in two places: the WEIGHT statement and the TOTCNT statement. The 10% subsample contains the same hospitals as the full NIS sample, but has only 10% of their discharges rather than the 100% contained in the NIS.

WEIGHT

When using the 10% sample to get weighted estimates, multiply the variable DISCWT_U by 10, and specify the resulting variable on the WEIGHT statement.

TOTCNT

The TOTCNT statement must be modified because SUDAAN counts the observations in the sample at each level and uses the result instead of whatever is specified in the SAMCNT statement. For example, suppose an analyst wants to calculate mean LOS for AMI patients as above, but wants to use a 10% sample. Suppose the analyst specifies N_HOSP_F in the TOTCNT statement, as above, for the number of hospitals on the frame, but specifies TOTDSCHG for the total number of discharges for each hospital. The analyst would then specify S_HOSP_U for the number of sample hospitals in the stratum, along with a newly created variable that contained the number of discharges in the 10% sample from each hospital, .10*TOTDSCHG.

However, as mentioned above, SUDAAN will double check the counts of the variables specified in the SAMCNT statement. When it counts the hospitals, it will determine the correct number. But when it counts the discharges on the analysis file, it will find far fewer than 10% of each hospital's actual discharges because the analysis file contains only AMIs. SUDAAN will use the count of AMI discharges instead of all discharges in the 10% sample, which will cause it to underestimate the finite correction f_{hi}, thus overstating the variance.

There are at least two ways to modify the TOTCNT statement to avoid this problem when using the 10% sample:

1. The first is to specify _MINUS1_ instead of TOTDSCHG in the TOTCNT statement. This essentially tricks SUDAAN into calculating the variance without implementing the finite correction factor for this component, since specifying _MINUS1_ signals to SUDAAN that sampling at this level was with replacement. Since the finite correction factor for a 10% sample would be .9, calculating the variance this way will lead to a slight overestimate of the variance.

2. A more precise way would be to create a variable equal to the total number of AMIs in the hospital and use that in the TOTCNT statement instead of the total number of discharges in the hospital. This could be estimated as ten times the number of AMI discharges from the 10% sample or the actual number of AMI discharges from the full NIS sample file.

When SUDAAN counts the discharges in the analysis file under either of these approaches and takes the ratio of that number to the variable specified in the TOTCNT statement, it will get about .1, the correct value for f_{hi}.

CALCULATING VALUES FOR SUBSETS OF THE NIS

The SUBGROUP and LEVELS statements from SUDAAN can be used to calculate values for subsets of the entire population using either the full NIS or any NIS subsample. In either case, the entire data set is used. As an example, suppose an analyst desires values for male and female discharges. The analyst can use a variable SEX, where SEX = 1, if the discharge is a male, and SEX = 2, if female. To produce the proper code, the analyst modifies the code given in Section III, if using the full NIS, or Section IV, if using a NIS subsample. The modifications are:

1. Remove the SAMCNT statement (SUDAAN counts the cases) and

2. Add the statements:

   ```
   SUBGROUP SEX;
   LEVELS 2;
   ```

   to designate the partitioning variable and the numbers of partitions.

One can calculate values for more than two subsets in a single step. For example, if an analyst desires m partitions for a variable, the analyst creates variable P with values 1, 2 ...m, which partition the data set into the desired subsets (SUDAAN specifies that only the integers 1, 2, ...m can be used to create the m subsets). The statements:

```
SUBGROUP P;
LEVELS m;
```

are used to denote the partitioning variable and the number of partitions.

More complex partitioning using crosses of multiple variables is also allowed. Details may be found in SUDAAN documentation.

COMPUTER RESOURCE REQUIREMENTS

SUDAAN is an expensive package to run on a mainframe, especially on samples as large as those generated by the NIS, so it may be wise to consider using the PC version of SUDAAN whenever possible. Although resource requirements will vary across applications, a program using SUDAAN to calculate 500 means and their variances for two variables from an analysis file that contained 625,000 observations took 10 minutes on a 100 Mhz Pentium. The program from the example above that calculated the mean length of stay and its standard error for AMI patients took about 5 minutes on the same machine. Files can be manipulated on the mainframe and then downloaded in either ASCII or SAS format to the PC, where they can be read by SUDAAN.

ENDNOTES

1. See Carlson, B. L., A. E. Johnson, and S. B. Cohen, 1993, "An Evaluation of the Use of Personal Computers for Variance Estimation with Complex Survey Data", *Journal of Official Statistics*, 9(4): 795-814 for additional information on these and other complex survey data packages.

2. These finite correction factors ($(1-f_h)$) and ($(1-f_{hi})$) reduce the variance as the sample becomes a larger portion of the frame. This is intuitively plausible since the variance would be zero if the entire frame were included in the sample.

3. SUDAAN may be purchased from the Research Triangle Institute, Research Triangle Park, NC. See Carlson *et al., op. cit.* for sources for other complex survey design packages.

4. The SUDAAN documentation states that the universe count should appear in the TOTCNT statement. That is because in many surveys, the frame *is* the universe. In the NIS the frame is not the universe of all hospitals U.S., as mentioned above, so the appropriate variable for the TOTCNT statement is the count of hospitals in the frame in each stratum.

5. Note that there are approximately 10 to 15 hospitals in each year of the NIS that provide less than a full year's worth of data. For those hospitals the analyst would not have all AMI discharges. However, the effect on the estimated variance is small enough to ignore.